A new Algorithm for Accelerating Pair-Wise Computations of Melting Temperature

Lars Kaderali [a] Alexander Schliep [a]

[a] *ZAIK/ZPR, Universität zu Köln, Weyertal 80, D-50937 Köln, Germany*

Key words: Probe Selection, DNA chips, Suffix Trees, dynamic programming, Thermodynamics, Hybridization

1 Introduction

Both medicine and biology need efficient diagnostic tests to measure tissue- or cell-specific expression of hereditary information. The availability of complete genome sequences will permit interesting questions to be asked and answered at the genome level rather than at the level of the individual gene. Unfortunately, traditional tools are no longer capable to efficiently support the size of assays required for such tasks. It is thus not surprising that DNA chips, which allow to parallelize hybridization — the weak binding of two strands of DNA — experiments, are widely used in biological research.

Given a set of genomic sequences, the target sequences, we have to find at least one probe for each target sequence in the set. These probes will then be attached to the chip surface. The problem is that each probe on the chip should hybridize only to the intended target, and not to any other sequence in the target set, i.e., a probe must have a high specificity in detecting the target. Also, all probes must work under the same hybridization conditions, most importantly, at the same temperature. The problem can be formalized as follows:

Given n target sequences $t_1, t_2, ..., t_n$, find a temperature T and n probe sequences $p_1, p_2, ..., p_n$ such that

$$T_M(p_i, t_i) - \epsilon > T > T_M(p_i, t_k) + \epsilon \tag{1}$$

Email address: kaderali@zpr.uni-koeln.de (Lars Kaderali).

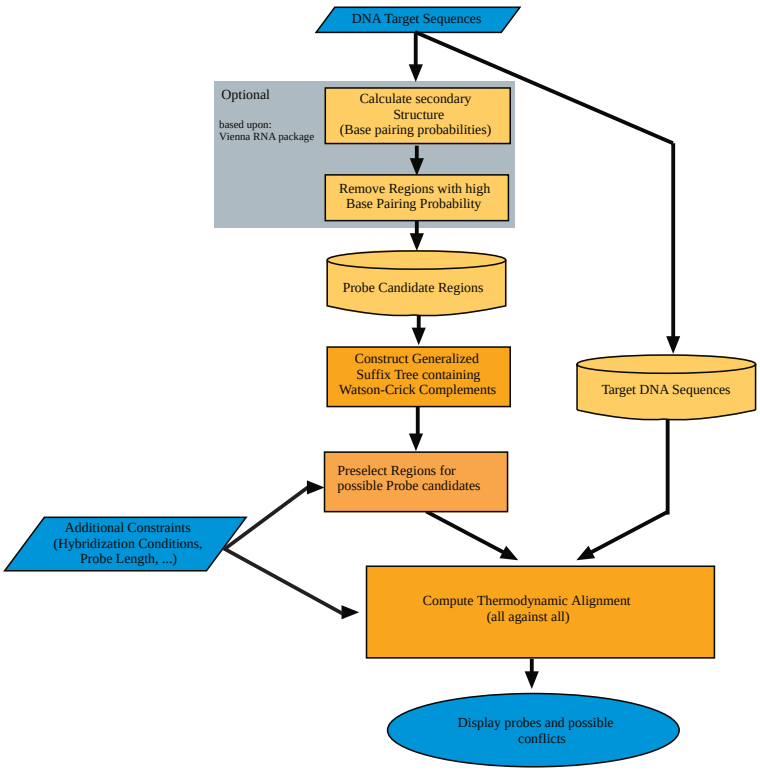

Fig. 1. Method overview for probe selection algorithm:Given the DNA target sequences, our goal is to exclude infeasible probe candidates as early as possible. Infeasible probes are probes that are too short or too long, occur more than once in different probes, or do not fulfill other relevant criteria. Thermodynamic considerations will only be made for the remaining probes, which reduces the problem's complexity considerably.

for all $k \neq i, i = 1, ..., n$, where $T_M(x, y)$ is the temperature below which the two strands x and y are bound, and above which they denature. T denotes the temperature at which the chip experiment should be carried out. ϵ is an additional temperature margin, compensating for example for model errors and imprecisions.

Computing Melting Temperatures

Interactions between bases in nucleic acids are of two kinds [1]:

- *Base pairing* in the plane of the bases due to hydrogen bonding between base pairs in the two opposing strands.
- *Base stacking* perpendicular to the plane of the bases due to London dispersion forces and hydrophobic effects.

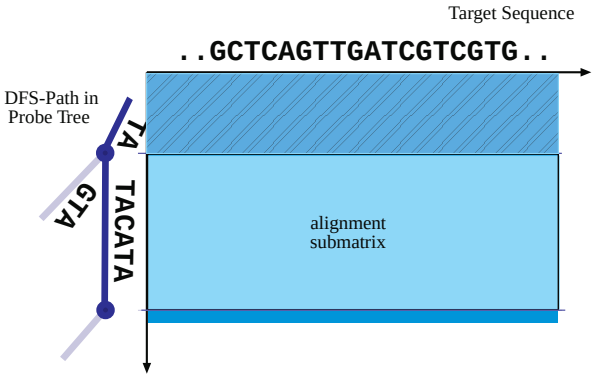

Fig. 2. Alignment of the target sequence with probe $..TATACATA...$ Note that if the alignment of the sequence TAGTA and the same target has been computed before, the upper part of the dynamic programming table can be reused.

Both quantum chemical calculations and thermodynamic measurements suggest that base pairing contributions to total energy depend exclusively on base pair composition, while stacking contributions depend on base pair composition and base sequence along the chain. Obviously, models based solely on base composition neglect stacking contributions, and yield less precise results [3]. As the major contribution to the overall stabilizing energy of nucleic acid structures comes from short-range interactions, we assume that the stability of a base pair depends only on the identity of its immediate up- and downstream neighbors [1]. This assumption leads to the Nearest Neighbor (NN) Model.

Algorithm

To select optimum probes, melting temperatures between the complements of all substrings of all target sequences (as the probe candidates) and all targets have to be computed. Therefore, an alignment between the probe and the target sequence is required, and the alignment resulting in the highest melting temperature T_M is desired. As it is possible that the probe-target-duplex contains secondary structure, all combinations containing loops et cetera should be considered as well. The running-time can be reduced considerably, by excluding bad probe candidates as early as possible, as then no alignment has to be computed for that candidate. Also, as prefixes of substrings are likely to occur several times in different substrings, avoiding to recompute entropy and enthalpy values for duplexes involving such prefixes will reduce runtime further.

The biggest improvement in running-time comes from the use of suffix [2] trees, which allows to re-use parts, cf. Fig. 2, of the alignment, or dynamic programming, matrix in the computation of the melting temperature. We will report

on the details of thermodynamic alignment algorithm, the use of generalized suffix trees and computational results obtained with our thermodynamic tree alignment algorithm for the probe selection problem.

References

[1] J. Cupal: *The Density of States of RNA Secondary Structures*, Diploma Thesis, Universität zu Wien, Formal- und Naturwissenschaftliche Fakultät, 1997.

[2] D. Gusfield: *Algorithms on Strings, Trees, and Sequences. Computer Science and Computational Biology*, Cambridge, 1997

[3] W. Rychlik, R. E. Rhoads: *A computer program for choosing optimal oligonucleotides for filter hybridization, sequencing and in vitro amplification of DNA*, Nucleic Acids Research 17, Number 21, 1989, p8543-8551